

# Limits of Performance of Quantitative Polymerase Chain Reaction Systems

Haris Vikalo, *Member, IEEE*, Babak Hassibi, *Member, IEEE*, and Arjang Hassibi, *Member, IEEE*

**Abstract**—Estimation of the DNA copy number in a given biological sample is an important problem in genomics. Quantitative polymerase chain reaction (qPCR) systems detect the target DNA molecules by amplifying their number through a series of thermal cycles and measuring the amount of created amplicons in each cycle. Ideally, the number of target molecules doubles at the end of each cycle. However, in practice, due to biochemical noise the efficiency of the qPCR reaction—defined as the fraction of the target molecules which are successfully copied during a cycle—is always less than 1. In this paper, we formulate the problem of the joint maximum-likelihood estimation of the qPCR efficiency and the initial DNA copy number. Then, we analytically determine the limits of performance of qPCR by deriving the Cramer–Rao lower bound on the mean-square estimation error. As indicated by simulation studies, the performance of the proposed estimator is superior compared to competing statistical approaches. The proposed approach is validated using experimental data.

**Index Terms**—Cramer–Rao lower bound, quantitative polymerase chain reaction.

## I. INTRODUCTION

**I**n *vitro* exponential amplification and quantification of specific sequences of nucleic acids (e.g., DNA, RNA) via the polymerase chain reaction (PCR) process [1], [2] is an essential part of many biotechnological procedures. Applications of PCR [3], [4] include sequence detection and genotyping, viral and bacterial detection, cancer diagnostics and genetic fingerprinting, DNA computing, to name just a few. Indeed, PCR-based protocols are widely used in nucleic acid research, and are one of the “gold standard” quantitative assaying techniques. Quantitative PCR, for instance, is the definitive benchmark for validation of gene expression data efficiently screened by DNA microarrays. The PCR technology, broadly defined, enables replication (or so-called amplification) of specific sequences of DNA *in vitro* by using an enzymatic replication process in each of its temperature-regulated cycles (typically, 30–40 of them). PCR derives its name from one of its key components—a DNA polymerase enzyme which synthesizes DNA fragments using primed nucleic acid templates. A PCR cycle, illustrated in Fig. 1, consists of three distinct

phases: denaturing, annealing, and extension. During denaturing, the sample is heated (typically, to temperatures above 90 °C) to break the hydrogen bonds between strands of the target DNA fragments, creating twice as many single-stranded fragments. Each of the single-stranded DNA fragments serves as a template during the second phase, annealing, when the sample is cooled down to temperatures typically between 40–72 °C. At such temperatures, primers—short sequences of nucleotides, designed to be exact complements to specific regions on templates—hybridize to the templates. There are two types of primers, one for each of the two types of the templates. The primers serve as initiation sites for the DNA polymerase enzyme activated in the last phase of a cycle, extension. The sample is heated to 72 °C maximizing the rate of extension while ensuring that the partially extended primers remain attached to the templates. Ideally, at the end of the extension phase, there are twice as many double-stranded target DNA fragments as there were at the beginning of the cycle. This generally results in an exponential growth of the number of the target nucleic acid sequences.

With PCR, it is possible to amplify a few copies of a DNA sequence by 6–9 orders of magnitude, generating billions (or more) of identical copies of the same sequence. The need for such a high-level molecular replication arises when the target DNA concentration is not sufficient for experimentation or reliable detection (e.g., in applications such as molecular diagnostics and forensics). However, practical issues affect the amplification process adversely and the efficiency of PCR—defined as the probability of generating a replica of each template molecule—is smaller than the desired one. Random nature of the underlying biochemical enzymatic process leads to variations in the PCR yield. Moreover, creation of nonspecific byproducts in the replication process further diminishes quality of the PCR product and the information that one can extract from the generated signal. The probabilistic nature of the replication process is addressed in [5]–[8], where various stochastic models have been proposed to describe the uncertainties of the process. In [9], the mutations-related effects that adversely affect the efficiency of PCR have been studied.

Quantification of the amplicons (DNA molecules obtained by making copies of the initial DNA template) is generally based on measuring the light intensity originating from the fluorescent reporter molecules (so-called probes) incorporated into the created amplicons. Hence, the measured signal is proportional to the number of successfully created copies of the template molecule. There are three types of commonly used probes: SYBR Green, hybridization probes, and TaqMan probes. SYBR Green I is a dye that binds to double-stranded DNA after which its fluorescence level increases significantly. However, it is nondis-

Manuscript received May 04, 2009. Current version published February 24, 2010. This work was supported in part by the National Science Foundation under Grant CCF-0845730.

H. Vikalo and A. Hassibi are with the University of Texas at Austin, Austin, TX 78712 USA (e-mail: hvikaloutexas.edu; hassibi@systems.caltech.edu).

B. Hassibi is with the California Institute of Technology, Pasadena, CA 91125 USA (e-mail: arjang@mail.utexas.edu).

Communicated by O. Milenkovic, Associate Guest Editor for the Special Issue on Molecular Biology and Neuroscience.

Color versions of Figures 1–4 in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIT.2009.2037088

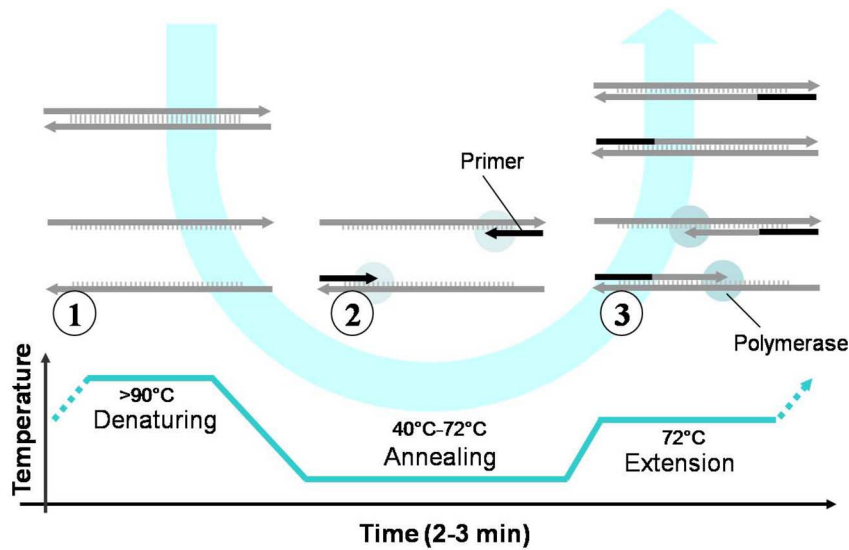


Fig. 1. Typical PCR cycle.

crimatory and just as likely to bind to nonspecific byproducts of PCR (e.g., primer-dimers) as it is to bind to the double-stranded target DNA. On the other hand, hybridization probes are specific to the target DNA sequence [10]. A hybridization probe consists of two short probe sequences, one labeled with fluorescein (a fluorescence resonance energy transfer donor) and the other with an acceptor dye. The two probe sequences are designed so that they hybridize next to each other on the target sequence; the colocation of the donor and the acceptor initiates energy transfer and, therefore, the change in their respective light intensities indicates successful replication of the target. TaqMan probes are also specific to the target sequence, and are designed so that they contain a fluorescein in the vicinity of a quenching dye. Once the TaqMan probe is attached to the template, a DNA polymerase separates fluorescein from a quenching dye, which results in an increased fluorescent signal.

In the real-time PCR, fluorescent signal is measured at the end of each temperature cycle. The measured light intensities comprise a reaction profile, usually plotted against the number of cycles. A typical reaction profile has three distinct phases: background phase, exponential growth phase, and saturation phase. During the first phase, the background noise originating from the unbound probes dominates the useful signal emanating from the probes attached to the templates. Although the fluorescent level of the unbound probes is much lower than the fluorescent level of the probes bounded to the double-stranded target DNA, the former significantly outnumber the latter during the first 15–20 cycles. The second phase starts when the signal from the PCR products rises sufficiently above the background noise. Typically, measurements collected during the exponential growth phase (also referred to as the log phase) are the only ones used to infer information about the original number of the DNA targets in the biological sample. The reason for imposing such a restriction is that the efficiency of PCR can be assumed constant in the first two phases, which makes the estimation tractable. In the third phase of PCR, however, the efficiency decreases rapidly as the reaction enters the plateau.

The ultimate goal of the quantitative real-time PCR is the estimation of the initial number of target molecules. In practice, this is commonly done by comparing a real-time PCR reaction profile with the reaction profiles of the so-called standard, where the latter are recorded for several initial concentrations of a target which has the same efficiency as the DNA target of interest. In recent work [11], [12], the reaction profile is first used to estimate the efficiency, which in turn is then employed to find an estimate of the initial number of the DNA target molecules.

The broad range of applications where the quantitative PCR is relied upon requires development of efficient estimation algorithms and the analysis of their performance limits. In this paper, we derive an approximate joint maximum-likelihood estimator of the PCR efficiency  $p$  and the number of initial target molecules  $x_0$ . Moreover, we find the Cramer–Rao lower bound on the minimum mean-square error of the estimated parameters, and demonstrate via simulations that the proposed estimator performs close to the Cramer–Rao bound. The paper is organized as follows. In Section II, we give the PCR model and derive the joint maximum-likelihood (ML) estimator of  $x_0$  and  $p$ . In Section III, we derive the Cramer–Rao lower bound on the minimum mean-square error of the estimator of  $x_0$ . Simulation results are presented in Section IV, and the experimental verification of the proposed technique in Section V. In Section VI, we summarize and discuss further implications of the results presented in the paper, outlining extensions to important applications of qPCR where the use of information-theoretic tools may prove indispensable.

## II. SYSTEM MODEL AND JOINT ML ESTIMATION OF $(x_0, p)$

Let  $x_0$  denote the initial number of target molecules which we want to estimate. We assume that the efficiency of replication during both the background phase and the exponential phase is constant, and denote it by  $p$ .<sup>1</sup> Furthermore, denote the number

<sup>1</sup>During the saturation phase, the efficiency drops as the reaction approaches plateau. For the sake of simplicity of the estimation procedure, we use only the measurements taken at the end of the cycles wherein the efficiency is constant.

of target molecules at the end of the  $n$ th cycle by  $x_n$ , and note that

$$x_n = (1 + p)x_{n-1} + \tilde{x}_n \quad (1)$$

where  $px_{n-1} + \tilde{x}_n \triangleq \xi_n$  is the number of amplicons that have been created in the  $n$ th cycle. Since the probability of replicating any of the  $x_{n-1}$  available amplicons in the  $n$ th cycle is  $p$ , it is easy to see that  $\xi_n$  is a binomial random variable with mean  $px_{n-1}$  and variance  $p(1-p)x_{n-1}$ . Recursion (1) describes a branching process, often used to model replication in biological systems [13]. It is not too difficult to show (see, e.g., [5]) that the mean of  $x_n$  in (1) is given by

$$\begin{aligned} E\{x_n\} &= E_{x_{n-1}} \{E_{x_n|x_{n-1}}\{(1+p)x_{n-1} + \tilde{x}_n\}\} \\ &= (1+p)E\{x_{n-1}\} = \dots = (1+p)^n x_0. \end{aligned} \quad (2)$$

Moreover, its variance can be found as

$$\begin{aligned} \sigma_n^2 &= E\{x_n - E\{x_n\}\}^2 \\ &= E_{x_{n-1}} \{E_{x_n|x_{n-1}}\{x_n^2\} - (E\{x_n\})^2\} \\ &= E\{E\{(1+p)^2 x_{n-1}^2 + 2(1+p)x_{n-1}\tilde{x}_n + \tilde{x}_n^2\} - (1+p)^2 x_0^2\} \\ &= E\{(1+p)^2 x_{n-1}^2 + p(1-p)x_{n-1}\} \\ &\quad - (1+p)^2 x_0^2 \end{aligned} \quad (3)$$

$$= \frac{1-p}{1+p} [(1+p)^{2n} - (1+p)^n] x_0 \quad (4)$$

where the transition from (3) to (4) can be shown by induction [5].

Imperfect instrumentation and other biochemistry independent sources create a noise which corrupts the measurements of  $x_n$ . We assume that the noise is additive Gaussian  $\mathcal{N}(0, \sigma_w^2)$ , and denote it by  $w_n$ . For the sake of notational simplicity, we set the transduction coefficient which maps the number of amplicons to the generated signal to  $k_t = 1$  ( $k_t \neq 1$  can be easily accounted for in the derived expressions). Thus, the quantity which is measured by is given by

$$z_n = x_n + w_n.$$

Let us denote the number of temperature cycles in the background phase of PCR by  $k$ . Therefore, the first measurement taken beyond the background noise level is  $z_{k+1}$ . Furthermore, denote the number of temperature cycles in the exponential phase by  $l$ . Hence, the last measurement taken before the efficiency starts rapidly deteriorating is  $z_{k+l}$ . Introduce a new variable,  $\mathbf{y}$ , defined as

$$\mathbf{y} = \begin{bmatrix} \frac{z_{k+1} - (1+p)^{k+1}x_0}{\sigma_{k+1}} \\ \frac{z_{k+2} - (1+p)^{k+2}x_0}{\sigma_{k+2}} \\ \vdots \\ \frac{z_{k+l} - (1+p)^{k+l}x_0}{\sigma_{k+l}} \end{bmatrix}.$$

Clearly, the variable  $\mathbf{y}$  is zero-mean. Finding the exact probability density function (pdf) of  $\mathbf{y}$  appears to be difficult. However, if  $x_0$  is sufficiently large, we can approximate the pdf of

$\mathbf{y}$  in the following way. Note that we can express  $\mathbf{y}$  as a sum of  $x_0$  identically distributed random variables

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \dots + \mathbf{y}_{x_0}$$

where each summand is given by

$$\mathbf{y}_i = \begin{bmatrix} \frac{x_{i,k+1} + w_{k+1}/x_0 - (1+p)^{k+1}}{\sigma_{k+1}} \\ \frac{x_{i,k+2} + w_{k+2}/x_0 - (1+p)^{k+2}}{\sigma_{k+2}} \\ \vdots \\ \frac{x_{i,k+l} + w_{k+l}/x_0 - (1+p)^{k+l}}{\sigma_{k+l}} \end{bmatrix}$$

where  $x_{i,j}$  denotes the number of created amplicons at the end of the  $j$ th cycle which had originated from the  $i$ th molecule in the original set of template molecules. Moreover,  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{x_0}$  are independent since they depend only on the number of amplicons generated in the first two phases of PCR, during which templates do not inhibit each other's replication. Therefore,  $\mathbf{y}$  can be represented as a sum of a large number of independent and identically distributed (i.i.d.) random variables. Hence, we can invoke the central limit theorem and argue that the distribution of  $\mathbf{y}$  may be approximated by a multivariate Gaussian distribution.

The  $(i, j)$ -entry of the  $l \times l$  covariance matrix of  $\mathbf{y}$ ,  $R$ , can be written as

$$R(i, j) = E\{y_i y_j\} = \frac{E\{z_{k+i} z_{k+j}\} - (1+p)^{2k+i+j} x_0^2}{\sigma_{k+i} \sigma_{k+j}} \quad (5)$$

where  $y_i$  and  $y_j$  denote the  $i$ th and the  $j$ th component of  $\mathbf{y}$ , respectively. Furthermore, note that

$$\begin{aligned} E\{z_{k+i} z_{k+j}\} &= E\{x_{k+i} x_{k+j}\} + \sigma_w^2 \delta_{i-j} \\ &= (1+p)^{j-i} E\{x_{k+i}^2\} + \sigma_w^2 \delta_{i-j} \\ &= (1+p)^{j-i} [\sigma_{k+i}^2 + (E\{x_{k+i}\})^2] + \sigma_w^2 \delta_{i-j} \\ &= (1+p)^{j-i} \sigma_{k+i}^2 + (1+p)^{2k+i+j} x_0^2 + \sigma_w^2 \delta_{i-j} \end{aligned} \quad (6)$$

where

$$\delta_{i-j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

and where we assumed (without a loss of generality) that  $i \leq j$ . Combining (5) and (6), we obtain

$$R(i, j) = (1+p)^{j-i} \frac{\sigma_{k+i}}{\sigma_{k+j}} + \frac{\sigma_w^2}{\sigma_{k+i} \sigma_{k+j}} \delta_{i-j}.$$

Now that we computed the covariance matrix  $R$ , the probability density function of  $\mathbf{y}$  can be approximated by the multi-variate Gaussian distribution

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{(2\pi)^{l/2} (\det R)^{1/2}} e^{-\frac{1}{2} \mathbf{y}^T R^{-1} \mathbf{y}}. \quad (7)$$

Note that  $f_{\mathbf{y}}(\mathbf{y})$  depends on  $x_0$  and  $p$  through both  $\mathbf{y}$  and  $R$ .

### A. Estimating the Parameters $x_0$ and $p$

The joint maximum-likelihood estimates of  $x_0$  and  $p$  can be found by solving the optimization problem

$$\max_{x_0, p} f_{\mathbf{y}}(\mathbf{y}),$$

or, equivalently, by solving the minimization

$$\min_{x_0, p} \{\mathbf{y}^T R^{-1} \mathbf{y} + \log \det R\}. \quad (8)$$

For convenience, we denote the objective function of the minimization (8) as  $\mathbf{y}^T R^{-1} \mathbf{y} + \log \det R \triangleq g(x_0, p)$ .

On the other hand, the traditional approach to the estimation of the initial population in a branching process (see, e.g., [14]) first focuses on finding the maximum-likelihood estimator of  $p$ ,

$$\hat{p} = \frac{z_{k+1} + \dots + z_{k+l}}{z_k + \dots + z_{k+l-1}} - 1. \quad (9)$$

Then, the above estimate  $\hat{p}$  is used to find  $x_0$  as

$$\hat{x}_0 = \frac{z_{k+l}}{(1 + \hat{p})^{k+l}}. \quad (10)$$

Note that the estimates  $\hat{p}$  and  $\hat{x}_0$  are computed using only the measurements taken in the exponential phase of the PCR process [5], [15].

Extensive simulation studies conducted over a wide range of parameters  $x_0$  and  $p$  indicate that the objective function (8), has a unique (i.e., global) minimum. Therefore, to solve (8), one can use, e.g., a simple gradient descent method initialized by  $\hat{x}_0$  and  $\hat{p}$  obtained from (10) and (9), respectively. In particular, let us collect the parameters that need to be estimated into a vector,  $\mathbf{c}^T = [x_0 \ p]^T$ . Starting from  $\mathbf{c}^T = [\hat{x}_0 \ \hat{p}]^T$ , we iteratively update (until a stopping criterion is satisfied)  $\mathbf{c} := \mathbf{c} + t\Delta\mathbf{c}$ , where

$$\Delta\mathbf{c} = -\nabla g(\mathbf{c}) = - \begin{bmatrix} \frac{\partial g}{\partial c_1} \\ \frac{\partial g}{\partial c_2} \end{bmatrix}.$$

The partial derivatives required for  $\Delta\mathbf{c}$  are computed as

$$\begin{aligned} \frac{\partial g}{\partial c_i} = & \left( \frac{\partial \mathbf{y}}{\partial c_i} \right)^T R^{-1} \mathbf{y} - \mathbf{y}^T R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \mathbf{y} \\ & + \mathbf{y}^T R^{-1} \frac{\partial \mathbf{y}}{\partial c_i} + \text{Tr} \left\{ -R^{-1} \frac{\partial R}{\partial c_i} \right\} \end{aligned}$$

for  $i = 1, 2$ , and the step size  $t$  is obtained via exact or backtracking line search. Alternatively, to improve the convergence speed of the algorithm, we may use, e.g., the Newton's method. [The explicit expressions for  $\partial \mathbf{y} / \partial c_i$  and  $\partial R / \partial c_i$  are provided in the next section, where they are computed as an intermediate step towards the calculation of the Cramer–Rao bound.]

### III. LIMITS OF PERFORMANCE OF QPCR: THE CRAMER–RAO BOUND

The qPCR systems are capable of detecting the initial number of target DNA molecules over a very broad dynamic range (several orders of magnitude). We focus on the scenario where no prior information about  $x_0$  is available, and treat  $x_0$  as an unknown parameter (or a random variable with a uniform distribution having a very large support). Given the extremely broad range of  $x_0$ , it is reasonable to extend the parameter space to a continuous line. This ensures that the regularity condition required for the existence of the Fisher information is satisfied; it also provides computational advantages compared to dealing with a discrete parameter space. The performance bounds in this section are computed assuming such a relaxation. [Note: if a prior information about  $x_0$  is available —e.g.,  $x_0$  is known to belong to a relatively confined interval—one may employ discrete parameter estimators and find the lower bounds on their mean-square error performance.]

The minimum mean-square error of any estimation procedure is lower bounded by the Cramer–Rao bound [16]. We derive the Cramer–Rao lower bound (CRLB) and use it to quantify the limits of achievable performance of DNA copy number estimation in quantitative PCR.

The Fisher information matrix,  $F$ , is given by the negative of the expected value of the Hessian matrix of  $\log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})$ , i.e.

$$F = -E_{\mathbf{y}} \{ \nabla_{\mathbf{c}} \nabla_{\mathbf{c}}^T \log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y}) \}$$

where  $\mathbf{c}^T = [x_0 \ p]^T$ . Therefore, the entries of the  $2 \times 2$  matrix  $F$  are given by

$$F_{ij} = -E_{\mathbf{y}} \left\{ \frac{\partial^2}{\partial c_i \partial c_j} \log p_{\mathbf{y}|\mathbf{c}}(\mathbf{y}) \right\}$$

where, for compactness of the notation,  $c_i$  and  $c_j$  denote the entries of  $\mathbf{c}$  (i.e.,  $c_1 = x_0, c_2 = p$ ). Assuming an unbiased estimator, the CRLB on the minimum mean-square error of estimating  $x_0$  is given by

$$E\{(\hat{x}_0 - x_0)^2\} \geq [F^{-1}]_{11} \quad (11)$$

where  $[F^{-1}]_{11}$  denotes the  $(1, 1)$ -entry of  $F^{-1}$ . Similarly, the CRLB on the minimum mean-square error of estimating  $p$  is

$$E\{(\hat{p} - p)^2\} \geq [F^{-1}]_{22} \quad (12)$$

where  $[F^{-1}]_{22}$  denotes the  $(2, 2)$ -entry of  $F^{-1}$ .

Let us denote  $L_1(\mathbf{c}) = \log \det R$  and  $L_2(\mathbf{c}) = \mathbf{y}^T R^{-1} \mathbf{y}$ , so that we can write

$$L(\mathbf{c}) = \log(p_{\mathbf{y}|\mathbf{c}}(\mathbf{y})) = -\log(2\pi) - \frac{1}{2}L_1(\mathbf{c}) - \frac{1}{2}L_2(\mathbf{c}).$$

Therefore, the Fisher information matrix can be written as

$$F = \frac{1}{2}E_{\mathbf{y}} \{ \nabla_{\mathbf{c}} \nabla_{\mathbf{c}}^T L_1(\mathbf{c}) \} + \frac{1}{2}E_{\mathbf{y}} \{ \nabla_{\mathbf{c}} \nabla_{\mathbf{c}}^T L_2(\mathbf{c}) \}.$$

It is easy to find components of  $\nabla_{\mathbf{c}} \nabla_{\mathbf{c}}^T L_1(\mathbf{c})$

$$\frac{\partial^2}{\partial c_i \partial c_j} L_1(\mathbf{c}) = \text{Tr} \left\{ -R^{-1} \frac{\partial R}{\partial c_j} R^{-1} \frac{\partial R}{\partial c_i} + R^{-1} \frac{\partial^2 R}{\partial c_j \partial c_i} \right\} \quad (13)$$

where  $\text{Tr} \{ \cdot \}$  denotes the trace operation over its argument, and where

$$\frac{\partial R(i, j)}{\partial c_1} = -\frac{1}{x_0} \frac{\sigma_{\mathbf{w}}^2}{\sigma_{k+i} \sigma_{k+j}} \delta_{i-j}$$

$$\begin{aligned} \frac{\partial R(i, j)}{\partial c_2} &= (1+p)^{j-i-1} \frac{\sigma_{k+i}}{\sigma_{k+j}} (j-i) \\ &+ (1+p)^{j-i} \left( \frac{\sigma'_{p,k+i}}{\sigma_{k+i}} - \frac{\sigma'_{p,k+j}}{\sigma_{k+j}} \right) \\ &- \frac{\sigma_{\mathbf{w}}^2 \left( \frac{\sigma'_{p,k+i}}{\sigma_{k+i}} + \frac{\sigma'_{p,k+j}}{\sigma_{k+j}} \right)}{\sigma_{k+i}\sigma_{k+j}} \delta_{i-j} \end{aligned}$$

and

$$\frac{\partial^2 R(i, j)}{\partial c_1 \partial c_2} = \frac{1}{x_0} \frac{\sigma_{\mathbf{w}}^2}{\sigma_{k+i}\sigma_{k+j}} \left( \frac{\sigma'_{p,k+i}}{\sigma_{k+i}} + \frac{\sigma'_{p,k+j}}{\sigma_{k+j}} \right) \delta_{i-j}$$

where  $\sigma'_{p,j}$  and  $\sigma''_{p,j}$  denote the first and the second derivative of  $\sigma_j$  with respect to  $p$ , respectively.

Finding components of  $\nabla_c \nabla_c^T L_2(\mathbf{c})$  is somewhat more involved. First, note that

$$\begin{aligned} \frac{\partial}{\partial c_i} \mathbf{y}^T R^{-1} \mathbf{y} &= \left( \frac{\partial \mathbf{y}}{\partial c_i} \right)^T R^{-1} \mathbf{y} - \mathbf{y}^T R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \mathbf{y} \\ &+ \mathbf{y}^T R^{-1} \frac{\partial \mathbf{y}}{\partial c_i}. \end{aligned}$$

The second derivatives, thus, are given by

$$\begin{aligned} &\frac{\partial^2}{\partial c_i \partial c_j} \mathbf{y}^T R^{-1} \mathbf{y} \\ &= 2 \left( \frac{\partial^2 \mathbf{y}}{\partial c_i \partial c_j} \right)^T R^{-1} \mathbf{y} - 2 \left( \frac{\partial \mathbf{y}}{\partial c_i} \right)^T R^{-1} \frac{\partial R}{\partial c_j} R^{-1} \mathbf{y} \\ &+ 2 \left( \frac{\partial \mathbf{y}}{\partial c_i} \right)^T R^{-1} \frac{\partial \mathbf{y}}{\partial c_j} + 2 \mathbf{y}^T R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \frac{\partial R}{\partial c_j} R^{-1} \mathbf{y} \\ &- 2 \mathbf{y}^T R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \frac{\partial \mathbf{y}}{\partial c_j} - \mathbf{y}^T R^{-1} \frac{\partial^2 R}{\partial c_i \partial c_j} R^{-1} \mathbf{y}. \quad (14) \end{aligned}$$

It is straightforward to show that the first and the second derivatives of  $\mathbf{y}$  with respect to  $c_1 = x_0$  are given by

$$\begin{aligned} \frac{\partial \mathbf{y}}{\partial c_1} &= \mathbf{a}_{11} + \mathbf{b}_{11} \circ \mathbf{y} \\ \frac{\partial^2 \mathbf{y}}{\partial c_1^2} &= \mathbf{a}_{12} + \mathbf{b}_{12} \circ \mathbf{y} \end{aligned} \quad (15)$$

where

$$\begin{aligned} \mathbf{a}_{11} &= - \begin{bmatrix} \frac{(1+p)^{k+1}}{\sigma_{k+1}} \\ \vdots \\ \frac{(1+p)^{k+l}}{\sigma_{k+l}} \end{bmatrix}, \mathbf{b}_{11} = -\frac{1}{2x_0} \mathbf{1}_l, \\ \mathbf{a}_{12} &= \begin{bmatrix} \frac{(1+p)^{k+1}}{x_0 \sigma_{k+1}} \\ \vdots \\ \frac{(1+p)^{k+l}}{x_0 \sigma_{k+l}} \end{bmatrix}, \mathbf{b}_{12} = \frac{3}{4x_0^2} \mathbf{1}_l \end{aligned}$$

where the  $l$ -dimensional vector  $\mathbf{1}_l$  is defined as  $\mathbf{1}_l = [1 \ 1 \ \dots \ 1]^T$ , and where  $d \circ e$  denotes the Hadamard (direct) product of  $d$  and  $e$ , i.e.

$$d \circ e = \begin{bmatrix} d_1 \\ \vdots \\ d_l \end{bmatrix} \circ \begin{bmatrix} e_1 \\ \vdots \\ e_l \end{bmatrix} = \begin{bmatrix} d_1 e_1 \\ \vdots \\ d_l e_l \end{bmatrix}.$$

Similarly

$$\frac{\partial \mathbf{y}}{\partial c_2} = \mathbf{a}_{21} + \mathbf{b}_{21} \circ \mathbf{y}, \quad \frac{\partial^2 \mathbf{y}}{\partial c_2^2} = \mathbf{a}_{22} + \mathbf{b}_{22} \circ \mathbf{y} \quad (16)$$

where

$$\mathbf{a}_{21} = - \begin{bmatrix} \frac{(k+1)(1+p)^k}{\sigma_{k+1}} \\ \vdots \\ \frac{(k+l)(1+p)^{k+l-1}}{\sigma_{k+l}} \end{bmatrix} x_0, \mathbf{b}_{21} = - \begin{bmatrix} \frac{\sigma'_{p,k+1}}{\sigma_{k+1}} \\ \vdots \\ \frac{\sigma'_{p,k+l}}{\sigma_{k+l}} \end{bmatrix}$$

and where

$$\begin{aligned} \mathbf{a}_{22} &= \begin{bmatrix} \frac{(k+1)(1+p)^{k-2}}{\sigma_{k+1}} \left( 2(1+p) \frac{\sigma'_{p,k+1}}{\sigma_{k+1}} - k \right) \\ \vdots \\ \frac{(k+l)(1+p)^{k+l-2}}{\sigma_{k+l}} \left( 2(1+p) \frac{\sigma'_{p,k+l}}{\sigma_{k+l}} - k-l+1 \right) \end{bmatrix} x_0 \\ \mathbf{b}_{22} &= \begin{bmatrix} 2 \left( \frac{\sigma'_{p,k+1}}{\sigma_{k+1}} \right)^2 - \frac{\sigma''_{p,k+1}}{\sigma_{k+1}} \\ \vdots \\ 2 \left( \frac{\sigma'_{p,k+l}}{\sigma_{k+l}} \right)^2 - \frac{\sigma''_{p,k+l}}{\sigma_{k+l}} \end{bmatrix} \end{aligned}$$

and where  $\sigma'_{p,j}$  denotes the first, and  $\sigma''_{p,j}$  denotes the second derivative of  $\sigma_j$  with respect to  $p$ , respectively. Finally, the mixed second derivative can be found as

$$\frac{\partial^2 \mathbf{y}}{\partial c_1 \partial c_2} = \mathbf{a}_{m2} + \mathbf{b}_{m2} \circ \mathbf{y} \quad (17)$$

where

$$\begin{aligned} \mathbf{a}_{m2} &= \begin{bmatrix} \frac{(1+p)^k}{\sigma_{k+1}} \left( \frac{\sigma'_{p,k+1}}{\sigma_{k+1}} (1+p) + \left( \frac{1}{2x_0} - 1 \right) (k+1) \right) \\ \vdots \\ \frac{(1+p)^{k+l-1}}{\sigma_{k+l}} \left( \frac{\sigma'_{p,k+l}}{\sigma_{k+l}} (1+p) + \left( \frac{1}{2x_0} - 1 \right) (k+l) \right) \end{bmatrix} \\ \mathbf{b}_{m2} &= \frac{1}{2x_0} \begin{bmatrix} \frac{\sigma'_{p,k+1}}{\sigma_{k+1}} \\ \vdots \\ \frac{\sigma'_{p,k+l}}{\sigma_{k+l}} \end{bmatrix}. \end{aligned}$$

Substituting (15)–(17) into (14), and noting that  $E\mathbf{y} = 0$  and  $\text{Tr}\{AB\} = \text{Tr}\{BA\}$ , it is straightforward to show that

$$\begin{aligned} E_{\mathbf{y}} \left\{ \frac{\partial^2}{\partial c_i \partial c_j} L_2(\mathbf{c}) \right\} &= 2(\mathbf{1}_l^T (\mathbf{b}_{i2} \delta_{i-j} + \mathbf{b}_{m2} \delta_{|i-j|-1})) \\ &- 2 \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_i} \text{diag}(\mathbf{b}_{j1}) \right\} \\ &- 2 \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_j} \text{diag}(\mathbf{b}_{i1}) \right\} + 2\mathbf{a}_{i1}^T R^{-1} \mathbf{a}_{j1} \\ &- \text{Tr} \left\{ R^{-1} \frac{\partial^2 R}{\partial c_i \partial c_j} \right\} \\ &+ 2 \text{Tr} \{ R^{-1} \text{diag}(\mathbf{b}_{j1}) R \text{diag}(\mathbf{b}_{i1}) \} \\ &+ 2 \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \frac{\partial R}{\partial c_j} \right\} \end{aligned} \quad (18)$$

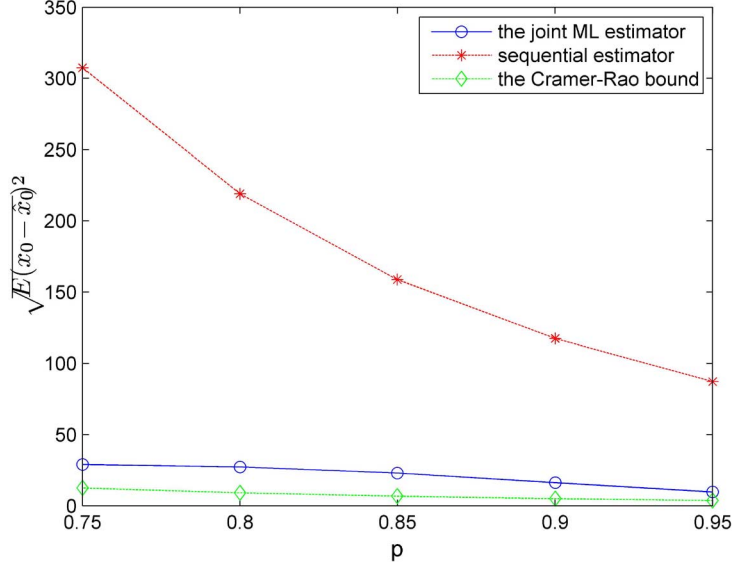


Fig. 2. Comparison of the estimators mean-square errors and the corresponding CRLB.

where  $\text{diag}(a)$  denotes a diagonal matrix with elements of vector  $a$  on the diagonal, and where  $\delta_k = 1$  if  $k = 0$  and 0 otherwise. On the other hand, from (13) it follows that

$$E_{\mathbf{y}} \left\{ \frac{\partial^2}{\partial c_i \partial c_j} L_1(\mathbf{c}) \right\} = -\text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_i} R^{-1} \frac{\partial R}{\partial c_j} \right\} + \text{Tr} \left\{ R^{-1} \frac{\partial^2 R}{\partial c_i \partial c_j} \right\}. \quad (19)$$

Combining (18) and (19), we finally arrive to

$$\begin{aligned} F_{ij} &= \frac{1}{2} E_{\mathbf{y}} \left\{ \frac{\partial^2}{\partial c_i \partial c_j} L_1(\mathbf{c}) \right\} + \frac{1}{2} E_{\mathbf{y}} \left\{ \frac{\partial^2}{\partial c_i \partial c_j} L_2(\mathbf{c}) \right\} \\ &= \mathbf{1}_l^T (\mathbf{b}_{i2} \delta_{i-j} + \mathbf{b}_{m2} \delta_{|i-j|-1}) \\ &\quad - \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_i} \text{diag}(\mathbf{b}_{j1}) \right\} \\ &\quad - \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_j} \text{diag}(\mathbf{b}_{i1}) \right\} + \mathbf{a}_{i1}^T R^{-1} \mathbf{a}_{j1} \\ &\quad + \text{Tr} \left\{ R^{-1} \text{diag}(\mathbf{b}_{j1}) R \text{diag}(\mathbf{b}_{i1}) \right\} \\ &\quad + \frac{1}{2} \text{Tr} \left\{ R^{-1} \frac{\partial R}{\partial c_j} R^{-1} \frac{\partial R}{\partial c_i} \right\}. \end{aligned} \quad (20)$$

We refrain from giving explicit expressions for the derivatives  $\sigma'_{p,j}$  and  $\sigma''_{p,j}$  and from substituting them in (18)–(20)—although trivial to obtain, these expressions are fairly cumbersome to write. Instead, we use a symbolic math manipulation package (in particular, *Mathematica*) to efficiently compute  $F_{ij}$  in (20) for a given set of parameters  $(\sigma_{\mathbf{w}}^2, k, l, x_0, p)$ , and thus obtain the CRLBs in (11) and (12).

#### IV. SIMULATION RESULTS

In Fig. 2, we compare the mean-square error of the estimate of  $x_0$  computed by (8) and that of (10), and compare them with the corresponding Cramer–Rao lower bound. A polymerase chain reaction is simulated as a branching process with  $x_0 = 1000$ ,

while the variance of the noise in the exponential phase is assumed to be 1/100 of the measured signal intensity. The estimation is based on only two data points (i.e., we assume  $l = 2$  measurements in the exponential case).

The proposed joint ML estimator (8) outperforms the estimator (10) for all considered values of  $p$ . Furthermore, the mean-square error of the joint ML estimator almost achieves the computed Cramer–Rao lower bound. On a related note, under fairly general assumptions an ML estimator is asymptotically efficient (i.e., should achieve the Cramer–Rao bound). In Fig. 2, the slight discrepancy is likely caused by the approximation of the true distribution of  $\mathbf{y}$  by a Gaussian.

#### V. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed algorithm on data obtained in a set of qPCR experiments. For a template DNA, we used pUC57 DNA Plasmid (essentially a circular DNA strand), a commonly used plasmid cloning vector in *E. coli*. The plasmid is isolated from *E. coli* by ion exchange chromatography. The molecule is a double-stranded circle and is illustrated in Fig. 3.

The region of amplification in this experiment is 2710 base pairs (bp) in length, and the two 25-bp-long primers used in the PCR have melting temperatures of 62.9 °C. In this experiment, we used 40 PCR identical cycles with temperatures of 60 °C degrees for annealing and extension phase, and 95 °C for the denaturing phase. Real-time detection of generated amplicons was carried out using SYBR Green fluorescent dye.

In Fig. 4, three PCR amplification curves are shown. Clearly, the signal initially has an exponential growth initially, but after a number of cycles the efficiency drops and the process becomes saturated. The amplification plots in this figure correspond to three different initial amounts of the DNA template (1 pg, 5 pg, and 10 pg). We employed both (8) and (10) to estimate the initial target amounts from the acquired PCR amplification data. The consecutive ratios of the estimated template amounts in the experiments are 5.27 and 12.51 for the algorithm (8), and 5.38

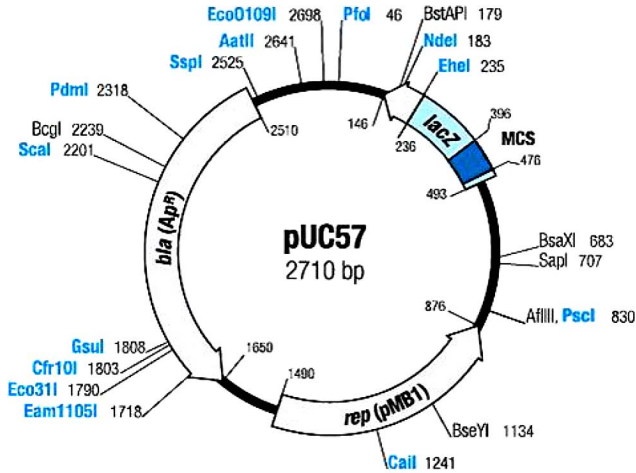


Fig. 3. pUC57 DNA Plasmid used as a template in qPCR experiments.

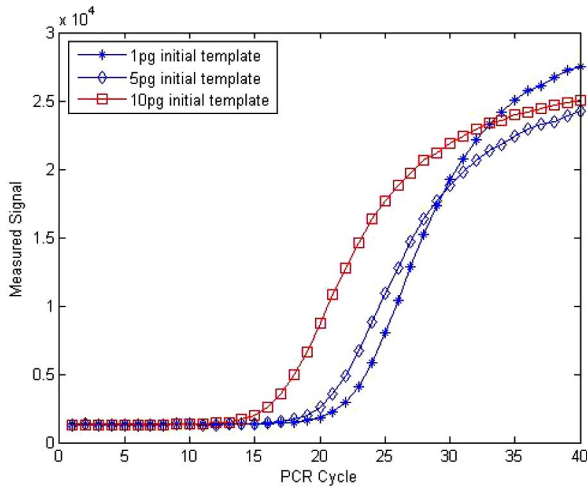


Fig. 4. QPCR signal for three different initial amounts of template (1 pg, 5 pg, and 10 pg).

and 12.55 for the algorithm (10) (the true values of the ratios are 5 pg/1 pg = 5 and 10 pg/1 pg = 10). The experimental results suggest applicability of the proposed algorithm to the initial template estimation in practical quantitative PCR systems.

## VI. SUMMARY AND DISCUSSION

In this paper, we studied the problem of joint estimation of the efficiency and the initial DNA copy number in quantitative polymerase chain reaction systems. We proposed a simple approximate model of the system, stated the maximum-likelihood estimation problem, and described a numerical approach to finding its solution. Then, we derived the corresponding Cramer–Rao lower bound on the mean-square estimation error, providing an analytical expression for the ultimate limit of the qPCR performance. As indicated by the simulation studies, the performance of the proposed estimator is superior with respect to the competing statistical approaches. Finally, we validated the proposed approach using experimental data.

In many applications, detection of the targets of interest is rendered difficult by the presence of strong biological interferers, i.e., by the presence of a large number of DNA fragments whose sequence is partly similar to that of the target template. This is very common in the situations where the sample originates from an organism, tissue, or a complex biological system. All molecular diagnostic applications essentially fall into this category. In infectious disease detection, for instance, it is typically required to amplify and detect less than 100 DNA strands of a pathogen in the presence of millions of copies of the host DNA. Those DNA fragments that bear similarity to the target template may also be amplified, albeit with lower probability (i.e., lower PCR efficiency). Nevertheless, since the initial copy number of such fragments may be several orders of magnitude larger than the initial copy number of the target template, the amplicons originating from them significantly contribute to the overall acquired signal. The goal of qPCR then becomes that of extracting the desired information about the target template from the measured composite signal. There are two parts to this problem: 1) determining the number of sources contributing to the total signal (i.e., the number of different DNA fragments—including the target template—that are being amplified) and 2) estimating the parameters associated with each of the sources. Detecting a number of sources from a composite signal is a well-studied problem in many fields. Traditionally, the number of sources is often detected using information-theoretic criteria such as the Akaike information criterion (AIC) or the minimum description length (MDL) [17]. Estimation of the parameters associated with the signal sources in the present problem (i.e., the initial copy numbers of the target template and the interfering DNA fragments) can be performed using a variety of techniques, including—but not limited to—total least squares [18], ESPRIT algorithm [19], modified Prony’s method [20], etc. Clearly, further studies of this problem and the analysis of the performance of qPCR in the presence of strong interferers are of utmost importance, and the use of information-theoretic tools in those studies may prove indispensable.

On another note, design of primers is the key step in a PCR experiment. To this end, a number of primer design techniques have been proposed (see, e.g., [21] and the references therein). However, such techniques are typically focused on DNA pairing, and do not take into consideration actual performance of the qPCR system. The Cramer–Rao bounds provide a link between the performance and the qPCR efficiency, which is primer sequence dependent. This connection could potentially be exploited and incorporated into the primer design procedure, which may prove particularly useful in the design of primers for applications where interferers adversely affect performance of qPCR.

## REFERENCES

- [1] K. Mullis and F. Faloona, “Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction,” *Meth. Enzymol.*, vol. 155, pp. 335–350, 1987.
- [2] P. M. Holland, R. D. Abramson, R. Watson, and D. H. Gelfand, “Detection of specific polymerase chain reaction product by utilizing the 5 to 3 exonuclease activity of thermus aquaticus DNA polymerase,” in *Proc. Nat. Acad. Sci.*, 1991, vol. 88, pp. 7276–7280.
- [3] M. A. Innis, D. H. Gelfand, and J. J. Sninsky, *PCR Applications: Protocols for Functional Genomics*. New York: Academic, 1999.



- [4] M. J. McPherson and S. G. Moller, *PCR: The Basics From Background to Bench*. London, U.K.: BIOS Scientific Publishing, 2000.
- [5] G. Stolovitzky and G. Cecchi, "Efficiency of DNA replication in the polymerase chain reaction," in *Proc. Nat. Acad. Sci.*, Nov. 1996, vol. 93, pp. 12947–12952.
- [6] A. Hassibi, H. Kakavand, and T. H. Lee, "A stochastic model and simulation algorithm for polymerase chain reaction (PCR) systems," *GENSIPS*, 2004.
- [7] P. Jagers and F. Klebaner, "Random variation and concentration effects in PCR," *J. Theoret. Biol.*, vol. 224, no. 3, pp. 299–304, 2003.
- [8] D. Piau, "Mutation-replication statistics of PCR," *J. Computat. Biol.*, vol. 9, no. 6, pp. 831–847, 2002.
- [9] D. Wang, C. Zhao, R. Cheng, and F. Z. Sun, "Estimating the mutation rate during error-prone polymerase chain reaction," *J. Computat. Biol.*, vol. 7, pp. 143–158, 2000.
- [10] C. T. Wittwer, M. G. Herrmann, A. A. Moss, and R. P. Rasmussen, "Continuous fluorescence monitoring of rapid cycle DNA amplification," *BioTechniques*, vol. 22, pp. 130–138, 1997.
- [11] C. Jacob and J. Peccoud, "Estimation of the parameters of a branching process from migrating binomial observation," *Advances in Applied Probability*, vol. 30, pp. 948–967, 1998.
- [12] N. Lalam, C. Jacob, and P. Jagers, "Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency," *Adv. Appl. Probab.*, vol. 36, pp. 602–615, 2004.
- [13] M. Kimmel and D. E. Axelrod, *Branching Processes in Biology*. New York: Springer-Verlag, 2002.
- [14] J.-P. Dion, "Estimation of the mean and the initial probabilities of a branching process," *J. Appl. Probab.*, vol. 11, pp. 687–694, 1974.
- [15] P. Kainz, "The PCR plateau phase—Toward an understanding of its limitations," *Biochimica et Biophysica Acta*, vol. 1494, pp. 23–27, 2000.
- [16] H. Cramer, *Mathematical Models of Statistics*. Princeton, NJ: Princeton University Press, 1946.
- [17] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.
- [18] E. M. Dowling *et al.*, "Exponential parameter estimation in the presence of known components and noise," *IEEE Trans. Antennas Propag.*, vol. 42, no. 5, May 1994.
- [19] R. Roy and T. Kailath, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 37, no. 7, Jul. 1989.
- [20] M. R. Osborne and G. K. Smyth, "A modified Prony algorithm for fitting sums of exponential functions," *SIAM J. Sci. Statist. Comput.*, vol. 16, pp. 119–138, 1995.
- [21] Y. Cao *et al.*, "Information theory-based algorithm for in silico prediction of PCR products with whole genomic sequences as templates," *BMC Bioinform.*, vol. 6, no. 190, 2005, doi:10.1186/1471-2105-6-190..

**Haris Vikalo** (M'09) received the B.S. degree from the University of Zagreb, Croatia, in 1995, the M.S. degree from Lehigh University, PA, in 1997 and the Ph.D. degree from Stanford University, CA, in 2003, all in electrical engineering.

He held a short-term appointment at Bell Laboratories, Murray Hill, NJ, in summer 1999. From January 2003 to July 2003, he was a Postdoctoral Researcher, and from July 2003 to August 2007, he was an Associate Scientist at the California Institute of Technology. Since September 2007, he has been with the Department of Electrical and Computer Engineering, the University of Texas at Austin, where he is currently an Assistant Professor. He is a recipient of the 2009 National Science Foundation Career Award. His research interests include genomic signal processing, stochastic signal processing, communications, and algorithm complexity.

**Babak Hassibi** (M'08) was born in Tehran, Iran, in 1967. He received the B.S. degree from the University of Tehran in 1989 and the M.S. and Ph.D. degrees from Stanford University, CA, in 1993 and 1996, respectively, all in electrical engineering.

From October 1996 to October 1998, he was a Research Associate at the Information Systems Laboratory, Stanford University, and from November 1998 to December 2000 he was a Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, NJ. Since January 2001, he has been with the Department of Electrical Engineering at the California Institute of Technology, Pasadena, where he is currently a Professor. He has also held short-term appointments at Ricoh California Research Center, the Indian Institute of Science, and Linköping University, Sweden. His research interests include wireless communications, robust estimation and control, adaptive signal processing, and linear algebra. He is the coauthor of the books *Indefinite Quadratic Estimation and Control: A Unified Approach to  $H^2$  and  $H^\infty$  Theories* (New York: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ: Prentice Hall, 2000).

Dr. Hassibi is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck best paper award of the American Automatic Control Council, the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering, and the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE). He has been a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY Special Issue on "Space-Time Transmission, Reception, Coding, and Signal Processing" and an Associate Editor for Communications of the IEEE TRANSACTIONS ON INFORMATION THEORY.

**Arjang Hassibi** (S'01–M'09) received the B.S. degree with honors from the University of Tehran, Iran, in 1997 and the M.S. and Ph.D. degrees from Stanford University, CA, in 2001 and 2005, respectively, all in electrical engineering.

From March 2005 to August 2006, he was a Postdoctoral Scholar at the Department of Electrical Engineering at the California Institute of Technology, Pasadena. Since August 2006, he has been with the Department of Electrical and Computer Engineering of the University of Texas at Austin, where he is currently an Assistant Professor. He has also held research positions in Barcelona Design, Stanford Genome Technology Center, Panorama Research Institute, and Xagros Genomics, which he cofounded in 2001. His main research areas are biosensors and bioelectronics, integrated sensors, DNA and protein integrated microarrays, biomedical VLSI systems, biological assay modeling, noise spectroscopy, and molecular diagnostics.